National Park Service - Southwest Alaska Network
Inventory & Monitoring Program

Data Mining Summary Report
For
Southwest Alaska Network

July 2003

Dorothy C. Mortenson
Southwest Alaska Network
National Park Service
240 West 5th Avenue, #114
Anchorage, Alaska 99501

Wendy E. Bryden
Kenai Fjords National Park
National Park Service
PO Box 1727
Seward, Alaska 99664

Report Number: NPS/AKRSWAN/????-2003-001(TBA)
File Number: AKR-SWAN/??????(TBA)

Funding Source:
US National Park Service, Inventory & Monitoring Program

Recommended Citation:

Mortenson, Dorothy C., Wendy E. Bryden. July 2003. Data Mining Summary Report. Inventory and Monitoring Program, Southwest Alaska Network. USDI National Park Service. Anchorage, AK. 20 pg.

TOPIC(s):

Information Management, Data Mining, Data Management

Keywords:

Data mining, data management, information management, metadata, bibliography, naturebib, inventory, monitoring

Acronyms:

AKSO	Alaska Support Office (also referred to as AKRO – Alaska Regional Office)
ALAG	Alagnak Wild River
ANIA	Aniakchak National Monument & Preserve
ARLIS	Alaska Resource Library Information System
FGDC	Federal Geographic Data Committee
GIS	Geographic information systems
I&M	Inventory & Monitoring (Program)
KATM	Katmai National Park & Preserve
KEFJ	Kenai Fjords National Park
LACL	Lake Clark National Park & Preserve
mp	Metadata parcer software, checks for errors in metadata
NPS	National Park Service
RM	Resource Management
SWAN	Southwest Alaska Network
USGS	US Geological Survey
WASO	Washington Support Office

Contents

Appendix List

Background and Planning:

Plans:

Cheat sheets:

Reports:

Executive Summary

The Inventory and Monitoring (I&M) Program requires a thorough Data Mining Project. The purpose of the Data Mining Project was to find and catalog data and information relating to natural resources within the park or in the vicinity of the park in order to develop the monitoring plan.

The Data Mining Project consists primarily of two types of documentation: a bibliography and metadata. The bibliography documents formal and informal reports, articles, books, etc. Metadata information documents databases, geographic information system (GIS) data, spreadsheets, etc. Both of these documents will be searchable using the National Park Service, NPS Focus website (http://focus.nps.gov/).

In order to document the end products listed above, the products themselves needed to be organized. The Data Mining Project developed a natural resource directory structure based on park staff needs. It drafted an Information Management Plan for the parks to help with future projects and products. This Plan explains the flow of information throughout a project and addresses backups of the natural resource computer files, data security and maintenance, project organization, and hardcopy document management.

Legacy data with very little known information about them were also organized and documented, but with less strict standards. This consisted of creating readme.txt files and documenting what is known about the information, and creating a "parking lot" directory of past employees' projects. It also consisted of general documentation of hardcopy file folders of past employees.

Having a completed list of metadata records and a bibliography database are good products, but they are hard to digest and fully comprehend. A summary report of the data found was generated consisting of the title, date, data type, publisher and abstract. This summary is interactive and available on the SWAN website. A bibliography listing of each park was also generated into a report and is also available as a .pdf file on the SWAN website:
http://www.nature.nps.gov/im/units/nw01/index.htm
or can be searched through the NatureBIB website (password required):
http://www1.nature.nps.gov/im/apps/npbib/index.htm

For the benefit of other data managers and park staff and to document the steps used in this Project, a series of "Cheat Sheets" were developed to provide direct technical instructions to accomplish a particular task. These are included as appendices in the Data Mining Summary Report.

Kenai Fjords National Park and Lake Clark National Park and Preserve have gone through the initial Data Mining Project. A follow-up to the Information Management Plan and modifications to the bibliography database should happen within fiscal year 2004. Katmai National Park and Preserve also represents the park natural resource information for Alagnak Wild River and Aniakchak National Monument and Preserve. Data Mining is scheduled for Katmai later Fall 2003 when park staff will be available.

Introduction

The Inventory and Monitoring (I&M) Program requires a thorough Data Mining Project. This includes documenting what information has been collected about the parks such as reports or data. The purpose of documenting this information is to a) make the information discoverable via a search on a website or some other means, b) provide enough information so the end user can determine if the report or data is of interest, and c) ensure the appropriate use of the information by documenting any restrictions or the quality of the information.

The documentation falls into two categories: a bibliography and metadata. The bibliography documents formal and informal reports, articles, books, etc. Metadata information documents databases, geographic information system data, spreadsheets, etc. Both require very similar information, such as title, date, abstract, and keywords. They have their differences, however, so they remain two separate entities. For example, a bibliography may contain call numbers, where as a spreadsheet may contain the definition of each field. Both entities have standard fields required and optional.

Both of these sets of information can be searched from one place called a clearinghouse. In the case of the National Park Service, a website is available called NPFocus which can search either the bibliography, metadata or both. The purpose of the Data Mining Project are to populate these two sets of information so they may be used by researchers.

An additional goal of the Data Mining Project was to bring better data management practices to the parks and assist in data organization. Sifting through drawers and shelves of reports and computer files, one cannot help but to gain an understanding of the flow of information. It is not the I&M Program's responsibility to take on all data management, but the Program can assist the parks to develop a plan to accomplish this. We helped the parks accomplish this where possible and when time permitted.

This report summarizes the efforts of the last several months including revision of electronic directory structure for Kenai Fjords National Park (KEFJ) and Lake Clark National Park and Preserve (LACL); development of metadata for both parks, updates of National Park Service bibliography, NatureBIB, and; some revision on organization of hardcopy data at KEFJ.

Background and Planning

The Data Mining Project stems from Step 2 in the "Recommended Approach for Developing a Network Monitoring Program". The website is: www1.nature.nps.gov/im/monitor/approach.htm

2. Summarize existing data and understanding.
One of the most important steps in the process of developing a monitoring strategy is the task of identifying, summarizing, and evaluating existing information and understanding of park ecosystems. Much of this needs to be done before the scoping workshop is held.
To accomplish this task, it is anticipated that most networks will need to hire, assign or contract at least one or two full-time persons (e.g., a Monitoring Coordinator and data management specialist) and allow at least a year prior to the scoping workshop for this step to be accomplished.
This step will include a literature review, a review of the Resource Management Plan (RMP), General Management Plan (GMP), and other applicable plans for each park, and an inventory of existing datasets and other information on park ecosystems.
Superintendents and other park managers should be interviewed regarding the key management issues facing their park and the types of information they need from the monitoring program.
Current or historical monitoring of natural processes and resources in each park should be summarized, including data from monitoring of fire effects, T&E species, water quality, air quality, physical processes/changes, and other resources. Data sets and the sampling design used should be evaluated to determine whether the monitoring is meeting the needs of park managers and is providing reliable and credible data to help manage the park. Maps showing the locations where monitoring has occurred should be prepared.
Monitoring that is being conducted by neighboring agencies, partners, and related parks should be identified and summarized to help determine where comparable data sets and sampling protocols exist.
Where understanding exists regarding cause-effect relationships between environmental stressors and the park's natural resources, or where the linkages among ecosystem components are understood, draft conceptual models should be prepared to help summarize this understanding.

To simplify the above, the I&M Alaska Support Office, Central Alaska Network and the Southwest Alaska Network met to discuss the interpretation of Data Mining and how it applies to Alaska. Following is how the group defined Data Mining. For more information on this meeting, please review the minutes in

Appendix A. Background information in preparation for this meeting may be found in Appendix B.

Definition of Data Mining:
Find and catalog data and information relating to natural resources within the park or in the vicinity of the park in order to develop the monitoring plan.

The Purpose of Data Mining:
Evaluate and understand the parks based on existing information in order to develop the monitoring plan.
Evaluate for comparison to anticipated monitoring data
Opportunistically, demonstrate good cataloging practices for non-Inventory & Monitoring purposes
Opportunistically, develop a standard operating procedure for updates/maintenance of data and bibliographic cataloging.

The Heartland Network's "Data Mining Guidelines" written largely by Brent Frakes, Database Manager, was also a helpful resource. Please see Appendix C.

SWAN Work Plan:
SWAN developed a work plan to complete the Data Mining Project. This is an ongoing plan and is updated periodically. Please review Appendix D.

Electronic File Directory Clean-up

The underlying theme to the Data Mining Project is data organization: Knowing what you have and where it is. We do this by documenting the inventory of natural resource information either through a bibliography, if it is a document, or a metadata record, if it is data. Both the bibliography and metadata efforts will be further described later in this document.

To begin with some efficiency, we need to have our inventory of information sorted in some kind of order. If we were to document every single file without any kind of organization, our information would result with many unknowns, duplicates, and irrelevant records. We would also have a difficult time keeping our documentation updated, as files would be moved, deleted, or renamed with little regard.

The Southwest Alaska Network approached data mining by organizing the electronic information first. In hindsight, I believe this is one of the most important steps, if not *the* most important, in data mining. It is not the responsibility of the I&M Program to organize a park's electronic files. However, it is in the best interest of the park and the network to do so. It is a joint effort between the park and the network. The network can act as a facilitator, but it is ultimately the responsibility of the park to ensure the content is appropriately organized.

Steps to Electronic File Clean-up:
Ask questions; Be familiar with existing organization; Identify the problem areas
Discuss with the staff
Outline an ideal file structure
Create a backup and a time stamp file
Populate the new structure
Write standard operating procedure(s)
Follow-up

Step 1: Ask Questions/General Assessment

These types of questions were asked to get a general assessment of the initial situation:

What is the condition of the existing file structure and information?

- In general all SWAN parks needed some assistance in cleaning up files, either a complete file structure overhaul or some reassessment and tidying up.

What is the logic behind the structure?

- KEFJ used the same logic as their hardcopy file structure. This structure, however, resulted in many empty directories, buried directories, and was not seemingly intuitive to new staff. It did, however, provide a "home" for just about anything and there was some connection between the hardcopy file and the electronic file. This served to be the foundation of the newer structure.
- LACL resource management staff is relatively new and did not have a specific file structure in place.
- KATM has not been reviewed in detail yet. (Note KATM also houses ALAG and ANIA data).

What subjects contain lots of information (i.e., bear studies)? What subjects have very little information (i.e., invertebrates)?

Some topics, such as bear or salmon studies, have much more information than others. The file structure needs to accommodate this by bringing larger projects

and datasets to an upper directory. In the following example, the various bear studies are directly under \MammalsTerrestrial, as oppose to under \MammalsTerrestrial\Bears\HabitatStudy, etc.

Example:
\Resources
    \Data
        \Biological
            \MammalsTerrestrial
                \BearHabitatStudy
                \BearHumanStudy
                \BearSurveys

In considering a file structure, there are two approaches: "splitters" and "lumpers". If the structure is split too much, information gets buried. If the information is lumped together too much, directories become unwieldy. Somewhere in between is the balance and it may take a few attempts to reach this balance.

When was the last complete backup? Are backups reliable?

In all SWAN parks reliable backups are a significant issue. The park computer network staff is aware of these issues and is working to resolve them. In the meantime, a backup of the drives were specifically created before any file clean-up was started. The difference between a "catastrophic" backup (if the building burned down or the server fried) vs. a project backup (when you can logically remember completed stages of your project) was discussed with resource staff. Resource staff are encouraged to create their own project backups at particular milestones, such as after data entry, before project clean-up of files, and when the project is completed. Unfortunately, several of the resource staff do not have the appropriate hardware (CD writer) to do these types of backups. Requests have been made to upgrade these computers, not only to complete appropriate backups, but because these computers are also running Windows 98, have slow processors, and numerous other problems. The park computer network staff is aware of these issues and is working to resolve them as well.

Backup power was an issue for one of the parks. The server hosting all of the resource management data did not have a battery backup. This was quickly and easily resolved, but it should be noted it is important to check this detail.

What access problems and issues are there? For examples:
- Speed and access to certain drives?

- Sensitive information; is there some information that should be protected by law?
- Team projects; do staff need to share information with each other?

Each park has a list of "Drives" that automatically are assigned when a person logs onto their computer. The letters assigned are not necessarily consistent from park to park, but for sake of discussion, I will refer to the U, T, X, and W here. Please see Appendix F, Cheat Sheet: Mapped Drives for more details. In general park staff could not say with certainty which drive was used for what.

The U drive is the individual's own workspace that is on a server but is only available to that individual. It is much like a C drive, but is on a scheduled backup and contains only data and files (no programs). When I inquired about backups, I discovered that in reality, these drives were not being backed up due to some technical problems (which are being resolved).

The T drive is the team drive. In the parks, it generally is the whole park. There are subdirectories that separate the different divisions within the park. The T drive is located on a server within the park. It is only accessible to those in the park or who have special permission.

The X drive is the read-only GIS information, managed by the Regional Support Office and distributed to the parks. The X drive is located on a server or computer within the park. It is accessible to anyone within the park. All parks have the same general information on the X drive, and some park specific information as well.

The W drive is the Alaska Region drive, in which anyone within the Alaska Region can access. The W drive is located within the Alaska Regional Office and is, unfortunately, very slow and unrealistic to use for anything more than a general repository. Opening a dataset from this drive, for example, will bring even a computer located in the Regional Office to a halt and the computer will need to be restarted.

Sensitive information is collected for certain projects, such a telemetry information for radio collars, eagle nest locations, or endangered plant species locations. These need to be handled with the appropriate permissions and protection.

Sharing of information ranged within the park. In the case of LACL, the Resource Chief and the biologist were both involved in the same project, but in different capacities. The Resource Chief did more administrative and reporting

work, while the biologist did more data and analysis. This relationship was taken into consideration in the file structure to allow a combined project without stepping on each other's toes. This was also a similar relationship between the biologist and the biological technician.

Once a project was done, information needs to be brought forth to other levels, such as to the Superintendent, the GIS liaison, researchers from other agencies, future researchers of the Park Service, etc. How the flow of information was handled, from raw detailed datasets to "generally finished" to publication, needed consideration in the file structure.

Does staff know what they are supposed to do with the information they collect?
Not all staff understood where to find information, or knew what they were supposed to do with their own information. The idea is vaguely there, but not specific and not crystal clear. Some training and guidelines will help this, which is discussed later in this document.

Step 2: Discuss with staff

If the staff doesn't want an organized, resource management file structure, it isn't going to happen. Fortunately (with a great sigh of relief), everyone in the SWAN parks does. Most, if not all, are willing to do what is recommended, if only there was a recommendation.

Knowing how to best organize the resource management information needs input from the staff. There needs to be a balance between what a region-wide or agency-wide structure might be like and what is practical at the park level. Discussing how information is moved from the individual project "up the chain" to a national level needs to be considered. How will information be shared? How will it be archived?

Step 3: Outline an ideal file Structure

Initially we looked for an existing structure. There were some examples, such as the structure for GIS data or the NPS hardcopy file structure. This was added to the pool of things to take into consideration. Other considerations were:

- Limit the first two or three directory levels to a minimum (under 10 or 15).
- Name directories so they:
    - make sense,
    - sort logically (i.e., \MammalsMarine, \MammalsTerrestrial or \BearsHabitatStudy, \BearsHumanStudy), and
    - follow good naming structure (no special characters, spaces, etc.) that could be used in a pathname of a program without problems.

- Create a parking lot for things that might take longer to deal with, such as project information from past employees or unknown files.
- Create a read-only repository and a read/write directory for active or ongoing projects

Keep in mind it is not likely there will be one directory structure that will make everyone within the agency happy the first time around. This is an iterative process. Each resource management division within each park will have their own file structure and it will be different from the next division and from the next park. There has been some discussion to make one, catch all directory structure, but this will be discussed for quite some time. In the meantime, we will have organized our data, completed our data mining process, completed our standard operating procedures, etc. The good news is whatever structure does result in the end, we will have documented and organized files that can be reorganized with greater ease.

Step 4: Backup and Timestamp

Backups for the resource management files had not been completed in quite some time. Before starting anything, this needed to be done. In the case of KEFJ, networking from the resource management (RM) server to the backup servers was a complicated problem. Therefore the RM files were copied onto a "firebox", which is like a portable, large disk drive. The firebox was then physically carried and connected to the backup server and copied. In addition, groups of directories that could fit onto a CD were copied onto CD-ROMs. The information stayed on the firebox until the data was reorganized and in full production. The information was also on the original server, under a different directory (\Resources_Old) as read-only until the new structure is in full production.

In the case of LACL, the backups were also problematic. The needed hardware for backups was on order. Because a proper backup would not be in place before the data clean-up would take place, we used the "firebox" as a backup of the existing drive.

In the case of KATM, our timing was a bit off. Before we were able to meet and agree on a file structure, the spring and summer season was upon us. The necessary staff to complete this task is unavailable until the fall of 2003.

For comparison of file names, file sizes, etc. we used a freeware software package called Directory Lister v0.6 to create a timestamp file of all the files, pathanames and files sizes. Please see Appendix F: Cheat Sheet: Instructions for using Directory Server for more details. We considered creating a database of

tracking what files went to where, but found this would take too long and would not yield that much benefit. We looked for tools that might be able to do this automatically, but did not find one. Given that these parks have relatively small resource management divisions, staff felt comfortable with making the reorganization without this level of detail. Other networks with larger resources may need to give this further consideration.

Step 5: Populating the new structure

Timing is everything, particularly when dealing with seasonal projects and staff. The best time to take on a data clean-up project is in the winter. Permanent staff are back from the field and have wrapped up their field work. Having the file structure outline completed by fall lends itself to staff being able to populate the new structure and help with documentation over the winter months. Reorganization can also be completed before the spring season projects start, and, ideally, resource managers will have an infrastructure in place for the coming season projects.

We scheduled a two week period for park staff to organize their data into the new file structure. All resource staff of that particular park were made aware of the changes in advance and were advised to not make edits to either the old or the new file structures until the transition was completed. Two weeks was sufficient time to make the transition of the existing information. As time permitted, other files could be moved from individual C or U drives. Additional polishing off, such as writing README.TXT files for the main directories, could be done as time permitted.

For files that would take more time to sort out, such as work completed by past employees, a "parking lot" was set up. In this example it is called the \Users_old directory. Initial assessment of the information is written in a README.TXT file. Last names are added, if they are known (directory \Jane is changed to \Jane_Doe, for example.) As returning seasonals come back to the park, or the information is needed, these files can be cleaned up as time permits.

One issue we commonly encountered with KEFJ file clean-up involved dealing with older and unrecognized file extensions. Some information was stored in database and text software formats that no longer exist. Many of these databases were converted to current software applications. We created a short cheat-sheet of common files extensions that we encountered and how we dealt with them (Appendix F: Cheat Sheet: Tips on Opening Documents with Unknown File Extensions).

I should clarify not every single individual file is cleaned up. We weren't intending to finish projects that were incomplete or make everything perfect. The purpose is to sort, organize and tidy up a bit to make information discoverable and retrievable, and to establish a "home" for information. Likening it to other cleaning, we were cleaning enough to have friends over for a nice dinner at our house, not disinfecting a biological hazard laboratory. As time and interest warrants, further cleaning could be done. So in looking at the many bear studies, for example, you would find all these bear studies in the same general directory. For the individual projects, however, things might still be a little untidy. Future bear studies will hopefully be better organized, as a result of the Information Management Plan. An example of a draft plan can be found in Appendix E.

Step 6: Write up the Standard Operating Procedure or Information Management Plan

After the electronic file reorganization, we may result in an organized, resource management file structure, but how we got to this point should be documented. What do the parks want to do from here on out? What were the decisions made? What should new staff know? This leads us into the initial Information Management Plan and the Project Organizer guidelines for the individual parks. As examples, please see Appendix E.

Step 7: Follow-up

After the new file structure has been in place for a year and has lived through it's summer field season and fall wrap-up season, the structure should be revisited. Adjustments should be made as needed.

Metadata

Once the electronic data file structure was cleaned up in each park, we systematically reviewed files and created metadata for all Microsoft Excel spreadsheets and Access databases. Excel and Access are the two most commonly used file types use for data within the parks. Metadata was also created for some hardcopy data that did not have corresponding electronic data. We created metadata using the Spatial Metadata Management System 3.2 software (SMMS). Notes on how to use SMMS in found in Appendix G: Cheat Sheet: Notes on Using the SMMS Software and Completing FGDC metadata.

Metadata allows you to add theme keywords and place keywords. We added the following keywords to every record under the theme keyword list: NPS, and National Park Service, (inventory and/or monitoring when appropriate).

To ensure place names were spelled correctly and consistently, we created a place keyword thesaurus in SMMS for each park. First we used the US Geological Survey (USGS) place names point coverage in geographic information system (GIS) and clipped within a five mile buffer of the park boundaries. The resulting lists of place names and feature types were put into an Access database. This list was compared to USGS park maps and additional names were added. We exported this list, placed some necessary code (XML) at the beginning and the end, then imported it into SMMS. We added the following keywords to every record under the place keyword list: Alaska; four letter park code; park name. Instructions can be found in Appendix G: Cheat Sheet: Creating a Keyword List for SMMS. A Report of the list can be found in Appendix I: Reports: SWAN Gazetteer.

We exported metadata from SMMS and corrected it using the error checking options with the Metadata Parser (MP) provided by the Alaska Support Office Geographic Information Systems Office (AKSO GIS). Once the metadata was error free, they were put into two locations: One, a repository location in the \DataManagement directory, and; Two, in the individual folders where the data resides. In the later, the metadata record was renamed to the name of the file and given a .met extension. For example, the dataset baldeagle1998.mdb would have a metadata record named baldeagle1998.met. Instructions on using Metadata Parser are found in Appendix G: Cheat Sheet: Using Metadata Parcer (MP) to check SMMS Metadata for FGDC Compliant Errors.
The idea is that the metadata record should be found next to the actual data. If there are any changes to the data, it should be updated in the metadata record within the directory. Likewise if anyone were to find the data or want to copy the data, the metadata would be readily available to go with it. Periodically or at

the same time, the repository should be updated. The repository is a place where others can look for information without having to look in individual folders. It is also a place where metadata can be reviewed for consistency. Details of the flow of information are subject to change. The repository of information may be handled in the future through a website or through SMMS software. In any case, park staff should refer to the Information Management Plan for their park for updated procedures.

We also used MP to generate text and html metadata files from complete and corrected metadata. The HTML version is easier to read for content. These files were made available to park staff for review of the content of the information.

To present a "deliverable" regarding the metadata effort, we wanted to generate a report of a short list of the metadata, including title, date, data type, publisher, and abstract. We also wanted to make this list available on the internet, linking to the metadata records. In this way, all members of the Technical Committee and the park staff could view the metadata list at anytime at anyplace, and would not need any instructions on how to use it.

For geospatial metadata, the AKSO GIS Office has done an outstanding job of maintaining these records and making the metadata as well as the data available on the internet. Initially we thought we could tag onto this effort. The tabular data just inventoried in this exercise does not fit the same criteria as the geospatial data, however. Currently, the method used to extract the metadata involves using the Theme Manager in ArcView. These tabular data are not necessarily stand-alone datasets polished for the public and they do not generally have a GIS component to them. Because of their uniqueness, these metadata records and datasets didn't naturally fall into the same routine as the process used for GIS metadata.

We first looked to the WASO I&M Program for guidance. WASO is planning a revamping of the clearinghouse and will use NPFocus as the clearinghouse search tool. However, these tools are not yet in place. It is estimated these would be ready sometime in December of 2003. These tools do look promising and we would like to use this method along with the other I&M Networks.

Until WASO clearinghouse is ready for production, we needed a temporary solution. SMMS is an excellent tool for writing metadata, but it does not generate reports. SMMS has a clearinghouse option, for an additional fee, but this was too involved and expensive for a temporary solution.
Our temporary solution was as follows:
Export the records from SMMS to text format

Run MP, using a configuration file for formatting
Import the XML files into an Access 2002 database
Code the records to be either Biological, Cultural, Index, or Physical
Generate reports, save as .pdf and post on the website
Develop a website using ColdFusion to read the database
Website shows the metadata as a summary, dividing the records on to Biological, Cultural, Index, or Physical web pages.
Title links to the actual metadata record.

Using a sizable hammer, we were able to achieve our goal. If this becomes a longer term solution, the process should be revisited. It should be noted these metadata records are not available on any clearinghouse. They are only available on the website. This is acceptable for SWAN's intermediate needs. The Metadata Summary Report may be found in Appendix I.

Please note the following websites for more information:
Geospaital Metadata
http://www.nps.gov/akso/gis/index.htm
Kenai Fjords National Park: http://www.nps.gov/akso/gis/kefj/kefj.htm
Lake Clark National Park and Preserve: http://www.nps.gov/akso/gis/lacl/lacl.htm
Katmai National Park and Preserve and Alagnak Wild River: http://www.nps.gov/akso/gis/katm/katm.htm
Aniakchak National Monument and Preserve: http://www.nps.gov/akso/gis/ania/ania.htm

Tabular Metadata:
http://www.nature.nps.gov/im/units/nw01/MetaData/metahome.htm

FGDC Content Standards for Metadata:
http://www.fgdc.gov
http://www.fgdc.gov/metadata/meta_workbook.html

Bibliography

The National Park Service has two bibliographic databases that are maintained each serving a different purpose. One is called Voyager Library Catalog, the other is called NatureBIB. Voyager is used for cataloging a library's collection, where everything that is physically in the park library is recorded in this database. It does not include other information outside of the park library that might be about the park. It might also include information that has nothing to do with the park, such as a general reference book. To enter information into the Voyager, the off-the-shelf software called ProCite was used. ProCite files are sent to the NPS Librarian for update into Voyager. For normal NPS procedures, Voyager is used.

NatureBIB contains all bibliographic information about a park, whether the park holds a physical copy of it or not. It excludes reference material that are not specifically related to the park. NatureBIB combines many of the previously existing bibliographies such as DeerBIB, NRBIB, GRBIB, PaleoBIB, and others. To enter information into NatureBIB, a customized website data entry form or a customized downloadable Microsoft Access database can be used. For the Inventory and Monitoring Program, NatureBIB is used.

Both the Procite and NatureBIB have their own advantages and disadvantages, which can be compared in Appendix B. One of the issues we came across was determining if we were really creating a bibliography or a catalog. There is a need for both. The parks want to have just one database, one that could be maintained by a clerical staff. They do not want to maintain both. These two databases, however, do not use the same data standards and transferring records from one to the other is not that simple. ProCite would allow a person to classify the record as needed, say for catalog only or for SWAN Bibliography or for Water Quality Bibliography. The online version of NatureBIB does not give you that capability. The Access version of NatureBIB would allow you to customize a classifying table, but it would be lost as it was imported to the online version and would be lost with an updated download.

NatureBIB is still in its early stages of development. Understandably, resolving problems with this database are no small tasks. Updating and downloading of information have been slow, taking several months to a year to complete. All edits must be done online and there are no plans to change this. This has been problematic as one edit done online takes a minimum of 5 minutes to complete. The online interface is slow and cumbersome. There are plans, however, to improve the interface at an unspecified time in the future. We are hopeful this will make editing easier.

The information preloaded into NatureBIB are from a variety of sources, such as NRBIB and GRBIB. As such there are many duplicates. The duplicates, however, are not consistent and readily apparent and cannot be easily programmed. For example in comparison of two sources for the same document, the title may have been typed in slightly differently, different keywords were used, one might have an abstract while another has none, one may have been entered as a formal report, the other as an article. To clean-up these records, a desktop application and download of the data would be very helpful.

Resolving the duplicates issue gets a little more complicated. For one, the records in the online database (Oracle) does not appear to have a stable and unique ID. A person could not specifically say bibliography ID =1 should now be = 2. Bibliography ID = 2 will always be assigned to 2. So for example, if I were to receive a download of all SWAN records and make the corrections I wanted to make, there would be no way for WASO to replace those records (replace ID= 1 from me with ID=1 in Oracle). Information from WASO of why this is the case is sketchy, but I will trust there is a reason for it.

The other issue that makes resolving duplicates complicated is duplicates caused by the same document being entered by NRBIB and GRBIB. The task of editing the GRBIB records has been assigned to the WASO geologist. These records will be cleaned up independently of records from NRBIB, and will be done in the order of A-Z of the authors' last name. In other words, the same document could remain a duplicate and be cleaned up twice, in two slightly different ways. The advice has been not to attempt to clean-up the GRBIB records until WASO's tasks have been completed.

Given the current tools and advice and the slow turn around time to upload and download information, the effort to clean-up the records in NatureBIB outweighs the benefits. In the case of Alaska, most researchers will use ARLIS and consider it more reliable. It's not that having the NatureBIB database would not be helpful, as it would. It's that it will require so much more time and effort, that other significant data needs will suffer. In addition, it still does not meet the park's need for a single database to act as a catalog and a bibliography. Attempts to clean these records and to find more efficient ways of doing so will continue. It may be, however, NatureBIB needs more time to mature and the task is revisited in a year or so.

Despite the editing issues with NatureBIB, we did populate the database with new records. We searched file drawers and shelves in each park for resources

and bibliographic citations. Citations from KEFJ and LACL were checked to see if they had already been captured in the online version of NatureBIB. Citations which did not exist on line were added to a local version of NatureBIB (an Access based program). This information was then sent to WASO for uploading in the online version of NatureBIB.

Before the records were sent to WASO, the records were checked for consistency. NatureBIB interface does not have any editing capabilities. For example, say after you've entered a citation you noticed you misspelled someone's name. You cannot go back and edit that record to correct the spelling. To edit the records, you must open the Access tables directly to make your changes. To check for consistency and to tidy up the records, we developed an editing process of opening these tables. A description of this process can be found in Appendix H: Cheat sheet: NatureBIB Clean-up Procedures. It is recommended only those familiar with Access databases and NatureBIB use these procedures.

During the course of editing KEFJ's bibliography we found that the in-house version of the RM Procite database did not include many abstracts. Alaska Resource Library and Information Service (ARLIS) had made an effort several years ago to add abstracts to the database. We contacted them and they sent us a copy of bibliography files for KEFJ and other Alaska parks with the missing abstract files on CD. The CD contains this information in the 5.0 version of Procite however the RM version of the Procite database is still 3.4. This CD of information is located at Kenai Fjords National Park and the local files have not been updated.

Hardcopy Data

In the case of Lake Clark National Park & Preserve, the resource management division is relatively new and very small. Some of the work that has been done in the park are from various people passing through: students, researchers, volunteers, other agencies, etc. There may not be any final reports or any electronic files. There may be, however, a folder of valuable information that may date back many years. In some cases the current biologist on staff may not be familiar with the topic or have the time to be able to determine its value. In these instances, the resource management division has a few drawers of folders of "unknowns".

We attempted to capture this information as a one-time inventory. Any new projects will have a report and electronic files, and, hence, will have a bibliography or metadata record. This inventory captures information that could not be found electronically or in the bibliography.

In all probability, these files will be incomplete and may not yield great dividends. For this reason, we kept our inventory very simple, not using metadata or NatureBIB. We created an Access database with only a few fields. The purpose was to generate a list in a report with enough information to determine if it was worth pursuing further. If there were files of particular interest, we would complete a metadata or bibliography record, which ever was most appropriate.

The Access database contains fields for the file folder name, keywords, and abstract. The database documents if the information also exists in electronic format. Some file folders at the park were put into this format. Please see Appendix I: Reports: Clark NP&P Resource Management File Descriptions.

In Kenai Fjords National Park hard copy data is stored in the Resource Management file drawers using the old RM file structure. KEFJ plans to reorganize the hardcopy data so that administrative documents are kept in file drawers using the system instituted by Director's Orders 19. Project data will be pulled out of the file drawers and put in file boxes on designated shelves. Other reference materials will be pulled out of the main file drawers and placed on a shelf designated as the RM Reference Library. Both the project shelf and the reference shelf will be organized using the same format as the electronic file structure.

Conclusion

This document summarizes the work completed during the last several months for the Inventory & Monitoring Program, Southwest Alaska Network at Lake Clark National Park and Preserve and Kenai Fjords National Park. Additional efforts will be needed to complete the same work for Katmai National Park and Preserve, Aniakchak National Monument and Preserve and Alagnak Wild River.